

# Using the SRA Identifier Block

---

*20 Apr 2012 Draft D*

Features .....	3
Data Types .....	4
Data Structure .....	4
Compatibility .....	5
Semantics .....	5
Controlled Namespaces .....	5
Replacement tracking .....	5
Migration tracking .....	5
Persistence .....	6
Use Cases .....	6
Data Migration .....	6
Data Replacement .....	6
Data Equivalency .....	6
Examples .....	6
SRA document identifiers .....	6
SRA study reference .....	7
SRA Sample Reference .....	7
BioSample Reference .....	8
BioProject Reference .....	8
Replaced Record .....	8
Replacer Record .....	8
Elected Record .....	9
Successor Record .....	9
Submitter alternate identifiers .....	9
Submitter replaced identifiers .....	9
Commonly used external identifiers .....	10
Universally unique identifiers .....	10

## Overview

The purpose of the SRA Identifier block is to capture in one place all keys that are used as IDs. An ID can identify exactly one record within a context. A record may have multiple IDs. A record's ID must be unique within a context, and all objects in a context must have an ID. These properties do not hold for "names" or other monikers.

The ID base type, called IdentifierType, is a string, and it has attributes like ns, to indicate the namespace or context of the ID, is\_primary to indicate whether this ID is a primary key in a database, is\_active to indicate whether the ID is still active or has been deprecated or superceded, label to indicate whether and how to display the ID's tag. The ID base type is subclassed into types by business use: ID to indicate archive usage, XID to indicate external database usage, LOCAL to indicate a namespace constrained by the local document set (for example in a submission), and UUID to indicate universal unique identifier.

The XID object is not the same as the XRefType, which is a more general way to create linkages and relationships with objects in foreign (eg non-SRA) databases.

## XML Schema

The IdentifierType is defined in the SRA.common.xsd schema, please look in the following location(s):

- <file:///home/shumwaym/proj/SRA/sra/doc/SRA.common.xsd>
- [https://svn.ncbi.nlm.nih.gov/viewvc/toolkit/trunk/internal/trace\\_archives/sra/doc/SRA.common.xsd?view=markup](https://svn.ncbi.nlm.nih.gov/viewvc/toolkit/trunk/internal/trace_archives/sra/doc/SRA.common.xsd?view=markup)
- [http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA\\_1-4d/SRA.common.xsd?view=co](http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-4d/SRA.common.xsd?view=co)
- [http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA\\_1-4/SRA.common.xsd?view=co](http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-4/SRA.common.xsd?view=co)

Here is the relevant type code from SRA.common.xsd:

```

<xsd:complexType name="IdentifierNodeType">
  <xsd:annotation>
    <xsd:documentation>Abstract type for node in IDENTIFIER block.</xsd:documentation>
    <xsd:appinfo>
      <ann:Glossary/>
      <ann:Status current="Proposed"/>
    </xsd:appinfo>
  </xsd:annotation>
  <xsd:simpleContent>
    <xsd:extension base="xs:string">
      <xsd:attribute name="label" use="optional" type="xs:string">
        <xsd:annotation>
          <xsd:documentation>A string value that can be used as a display hint, or to
            qualify a non-SRA identifier.</xsd:documentation>
          <xsd:appinfo>
            <ann:Glossary/>
            <ann:Status current="Proposed"/>
          </xsd:appinfo>
        </xsd:annotation>
      </xsd:attribute>
    </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="AccessionType">
  <xsd:simpleContent>
    <xsd:extension base="com:IdentifierNodeType" /> </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="NameType">
  <xsd:simpleContent>
    <xsd:extension base="com:IdentifierNodeType">
      <xsd:attribute name="ns" use="optional" type="xs:string">
        <xsd:annotation>
          <xsd:documentation>A string value that constrains the domain of named
            identifiers (namespace).</xsd:documentation>
          <xsd:appinfo>
            <ann:Glossary/>
            <ann:Status current="Proposed"/>
          </xsd:appinfo>
        </xsd:annotation>
      </xsd:attribute>
    </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="UUIDType">
  <xsd:simpleContent>
    <xsd:extension base="com:IdentifierNodeType" /> </xsd:extension>
  </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="IdentifierType">
  <xsd:annotation>
    <xsd:documentation>Set of identifiers that can dereference the record. This element can
      specify a set of records each referenced by IdentifierType that can be the successor
      (replaces), the replacement set (split), that can be the replacer set (join), or can
      be the synonym (secondary accession).</xsd:documentation>
    <xsd:appinfo>
      <ann:Status current="Proposed"/>
    </xsd:appinfo>
  </xsd:annotation>
  <xsd:sequence>
    <xsd:element name="PID" type="com:AccessionType" minOccurs="1" maxOccurs="1" /> </xsd:element>
    <xsd:element name="SID" type="com:AccessionType" minOccurs="0" maxOccurs="unbounded" /> </xsd:element>
    <xsd:element name="XID" type="com:NameType" minOccurs="0" maxOccurs="unbounded" /> </xsd:element>
    <xsd:element name="LOCAL" type="com:NameType" minOccurs="0" maxOccurs="unbounded" /> </xsd:element>
    <xsd:element name="UUID" type="com:UUIDType" minOccurs="0" maxOccurs="1" /> </xsd:element>
  </xsd:sequence>
</xsd:complexType>

```

## Features

- Tracks archived assigned ids
- Tracks submitter assigned ids

- Tracks 3<sup>rd</sup> party assigned ids including catalog ids
- Can support secondary ids and aliases
- Can support UUIDs
- Tracks alternate or secondary ids assigned by different archives
- Tracks replacement of previously active records
- Tracks equivalence between records

## Data Types

The IdentifierNodeType abstract type extends xs:string with the following attributes:

- **label** – whether and how to display a tag string.

The IdentifierType is used as the identifier information for each XML document, for referencing other XML documents, and for referencing samples in the Sample POOL type.

Four concrete types subclass IdentifierType in order to suggest the business use of the ID:

**AccessionType** – A key in an NCBI primary database.

**NameType** – A key in an external database or namespace.

The following attributes are required:

- **ns** – Namespace (database) of the external name

Any number of external names may be present.

**UUIDType** – A key that is universally unique and requires no namespace.

## Data Structure

**PID** – A primary identifier, or key, in the SRA database (accession). Example: SRR000123. Exactly one primary identifier is required in every IDENTIFIER block. This value is equivalent to the document/@accession attribute.

**SID** – A foreign key in the SRA database (accession), or a defunct primary key in the SRA database. Example: PRJNA41443. Any number of secondary identifiers may be present.

**XID** – A key in an external database. Example: Coriell NA12878. Any number of external names may be present.

**LOCAL** – A key that resolves within the current set of documents. Exactly one local name must be present on submission. Local names are not needed for data download or exchange between archives. This value is equivalent to the (document/@alias, document/@center\_name) attribute tuple.

**UUID** – A key that is universally unique and needs no namespace. UUIDs are not used by the Archive but rather are provided as part of the SRA xml schema to serve downstream applications, including non-INSDC SRA mirrors.

## Compatibility

It is intended that the existing NameGroup and RefNameGroup types will continue to remain in use for backwards compatibility.

## Semantics

The IdentifierType is implemented by each SRA archive with additional business rules governing use of namespaces and scope of identifiers.

## Controlled Namespaces

Most namespaces are not interpreted and only apply to the current submission. The exceptions are listed here:

Reserved Namespaces (ID class)	Type	use
sra, era, dra	INSDC SRA database	Identify mirror records
biosample, bioproject, gds, gap, ..	NCBI Entrez databases	Identify NCBI dependency
ega, arrayexpress	EBI ENA database	Identify federated resource
coriell, atcc	sample vendor	Identify externally curated samples
TCGA, TARGET, EMMES_HMP, ...	Sample namespaces	Allow namespace/samplename identifiers
BI, BCM, BCCAGSC, WUGSC, JGI, ...	Center namespaces	Identify records within a center's namespace See the current centers list at : <a href="ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Centers/centers.tab">ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Centers/centers.tab</a>

## Replacement tracking

The IdentifierType can be used to name record(s) replaced (taken over) by the current record. The transitive closure of these replacing relations is a set of currently active records with replaced descendants. The converse relation (replaced by) can be computed from this forest so it is not tracked explicitly.

## Migration tracking

The replaced by relation would be tracked in the case where the record was replaced by a record in a new database (migrated), for example biosample or bioproject. Another case might be if a record was moved from one INSDC SRA to another and thereby received a replacement accession.

## Persistence

One goal of the IDENTIFIERS block is to document data migration, replacement, and equivalency relationships independently of the life cycle of the record, so that Archive users who form dependencies on a certain SRA record can always recover the relationship to other records even if it has been suppressed.

## Use Cases

### Data Migration

The ID block can be used to manage the transition of metadata from one record to another and provide a trackback mechanism to recover previous incarnations. This would include:

- Tracking a record in the archive (or prior to archiving) with a submitter supplied identifier.
- Tracking a record's identifier before and after a data migration.
- Tracking a record's identifier before and after a data consolidation.
- Tracking a changes in an identifier used for a dependency

### Data Replacement

The ID block can be used to indicate that the content has been replaced, and identify the previous record that represented the content. A run may have been mis-loaded due to errors in the original load process or a misrepresentation of the metadata that caused the data to be interpreted differently. If the result of the mis-load is an SRA archive image that is substantially different then the run's accession will be replaced. Another example is where duplicate runs have been discovered, and each run can be mapped to its duplicates although only one of them is retained in the archive.

### Data Equivalency

The ID block can be used to point to records that are equivalent and can be used interchangeably. An example is the BioProject and SRA study identifiers, which for a time will both be active identifiers of a study record (until migration from SRA study to BioProject is completed). Another example is where equivalent records have been discovered in multiple SRA instances. This would happen when a submitter has sent the same submission to both NCBI and EBI, for example. Over time, the INSDC may elect to retain one instance and suppress the other one, but the ID block can be used to maintain the equivalence relation.

## Examples

### SRA document identifiers

The document can contain IDENTIFIERS block in co-existence with existing NameGroup attribute group :

```
<RUN xmlns="" run_center="BI" run_date="2011-08-04T04:00:00Z" instrument_name="SL-HAC">
  <IDENTIFIERS>
    <PID>SRR354028</PID>
    <LOCAL ns="BI" >BI.PE.110804_SL-HAC_0370_BFCB02H8ACXX.6.UNMATCHED.srf</LOCAL>
  </IDENTIFIERS>
```

The document can contain IDENTIFIERS block in lieu of existing NameGroup attribute group:

```
<RUN>
  <IDENTIFIERS>
    <LOCAL ns="BI" >BI.PE.110804_SL-HAC_0370_BFCB02H8ACXX.6.UNMATCHED.srf</LOCAL>
    <PID>SRR354028</PID>
  </IDENTIFIERS>
```

This gives a migration path for adoption of Identifier block in place of the name group attributes group, or a method for reverse construction of the NameGroup attributes from the ID block.

## SRA study reference

Document dependency references to other documents can be encoded with or without the NameGroup attributes.

```
<STUDY_REF accession="SRP009022" refcenter="BI" refname="Ceratotherium_simum_simum_WGS">
  <IDENTIFIERS>
    <PID>SRP009022</PID>
    <SID>PRJNA74583</SID>
    <LOCAL ns="BI" >Ceratotherium_simum_simum_WGS</LOCAL>
  </IDENTIFIERS>
</STUDY_REF>

<STUDY_REF>
  <IDENTIFIERS>
    <PID>SRP009022</PID>
    <SID>PRJNA74583</SID>
    <LOCAL ns="BI" >Ceratotherium_simum_simum_WGS</LOCAL>
  </IDENTIFIERS>
</STUDY_REF>
```

## SRA Sample Reference

Where the sample reference is a simple reference, this can be represented with the Identifiers block:

```
<SAMPLE_DESCRIPTOR>
  <IDENTIFIERS>
    <PID>SRS293911</PID>
    <LOCAL ns="JGI">10908</LOCAL>
  </IDENTIFIERS>
</SAMPLE_DESCRIPTOR>
```

Where the sample is a pool and each member is identified with a sample id an Identifiers block can be used instead of (or in addition to) the NameGroup attribute group. This eliminates the need to always provide a default sample where the sample is not part of the pool (typically unidentified organism).

```
<SAMPLE_DESCRIPTOR>
  <POOL>
    <MEMBER proportion="1" member_name="tagged_908_TGCTCGAC">
      <READ_LABEL>barcode</READ_LABEL>
      <IDENTIFIERS>
        <PID >SRS267431</PID>
        <SID >SAMN739917</SID>
        <LOCAL ns="BI" >478560.5885.New Tech Library.SDZICR_KB13650</LOCAL>
      </IDENTIFIERS>
    </MEMBER>
  </POOL>
</SAMPLE_DESCRIPTOR>
```

## BioSample Reference

The successor BioSample record can be identified alongside the SRA sample accession, as in:

```
<SAMPLE_DESCRIPTOR>
  <IDENTIFIERS>
    <PID >SRS267431</PID>
    <SID >SAMN739917</SID>
    <LOCAL ns="BI">478560.5885.New Tech Library.SDZICR_KB13650</LOCAL>
  </IDENTIFIERS>
```

When the SRA Sample record becomes secondary, the PID/SID identifiers can indicate this:

```
<SAMPLE_DESCRIPTOR>
  <IDENTIFIERS>
    <PID >SAMN739917</PID>
    <SID >SRS267431</SID>
    <LOCAL ns="BI">478560.5885.New Tech Library.SDZICR_KB13650</LOCAL>
  </IDENTIFIERS>
```

When the SRA Sample record becomes defunct this fact will be reflected in the SRA database and livelist (not in the IDENTIFIER block).

## BioProject Reference

The successor BioProject record can be identified alongside the SRA Study accession, as in:

```
<STUDY_REF>
  <IDENTIFIERS>
    <PID >SRP010976</PID>
    <SID >PRJNA74601</SID>
    <LOCAL ns="JGI">10909</LOCAL>
  </IDENTIFIERS>
</STUDY_REF>
```

When the SRA Study record becomes secondary, the is\_active field can indicate this:

```
<STUDY_REF>
  <IDENTIFIERS>
    <PID>PRJNA74601</PID>
    <SID>SRP010976</SID>
    <LOCAL ns="JGI">10909</LOCAL>
  </IDENTIFIERS>
</STUDY_REF>
```

When the SRA Study record becomes defunct this fact will be reflected in the SRA database and livelist (not in the IDENTIFIER block).

## Replaced Record

The information that a record has been replaced is not indicated in the IDENTIFIERS block, but is tracked in the SRA database and livelist.

```
<RUN run_date="2008-11-24T23:08:44Z" instrument_name="GA-5">
  <IDENTIFIERS>
    <PID>SRR292241</PID>
  </IDENTIFIERS>
```

## Replacer Record

This example shows how a record, SRR390728, replaces a predecessor SRR292241:

```
<RUN run_date="2008-11-24T23:08:44Z" instrument_name="GA-5">
  <IDENTIFIERS>
    <PID>SRR390728</PID>
    <SID>SRR292241</SID>
  </IDENTIFIERS>
```

## Elected Record

This example shows how one record, SRR351940, has replaced 9 others (elected as successor), as in the use case where one run is selected for cSRA loading and the remaining runs are suppressed.

```
<RUN>
  <IDENTIFIERS>
    <PID>SRR351940</PID>
    <SID>SRR351941</SID>
    <SID>SRR351942</SID>
    <SID>SRR351943</SID>
    <SID>SRR351944</SID>
    <SID>SRR351945</SID>
    <SID>SRR351946</SID>
    <SID>SRR351947</SID>
    <SID>SRR351948</SID>
    <SID>SRR351949</SID>
  </IDENTIFIERS>
```

## Successor Record

This example shows how one record, SRR351940, has replaced another kind of record, analysis object SRZ019522.

```
<RUN>
  <IDENTIFIERS>
    <PID>SRR351940</PID>
    <SID>SRZ019522</SID>
  </IDENTIFIERS>
```

## Submitter alternate identifiers

Submitted records can retain their alternate identifiers and these can be treated as identifiers rather than attributes of the record. The label attribute calls out the display field.

```
<RUN center_name="BI" alias="70291ABXX110301.7.tagged_393.bam" run_center="BI" run_date="2011-03-01T05:00:00Z" instrument_name="SL-HBZ" accession="SRR404010">
  <IDENTIFIERS>
    <PID>SRR404010</PID>
    <LOCAL ns="BI">70291ABXX110301.7.tagged_393.bam</LOCAL>
    <LOCAL ns="BI" label="read group platform unit" >70291ABXX110301.7.CCAGTTAG</LOCAL>
  </IDENTIFIERS>
...
```

## Submitter replaced identifiers

Submitters can replace an identifier with a new one without disturbing the linkage to existing SRA accessions. However, the primary identifier must be supplied and the defunct identifier must be removed by an update submission.

existing...

```
<RUN alias="454_O.mykiss_GD3412001" accession="SRR090454" center_name="INRA">
  <IDENTIFIERS>
    <PID>SRR090454</PID>
    <LOCAL ns="INRA">454_O.mykiss_GD3412001</LOCAL>
  </IDENTIFIERS>
...
```

updated...

```
<RUN alias="454_O.mykiss_GD3412001" accession="SRR090454" center_name="INRA">
<IDENTIFIERS>
  <PID>SRR090454</PID>
  <LOCAL ns="INRA">454_O.mykiss_GB5RBPX02 </LOCAL>
</IDENTIFIERS>
```

...

## Commonly used external identifiers

In lieu of a local identifier, a submitter can use a supported external identifier. A good example is a cell line DNA isolate sample from one of the Coriell NA12878:

```
<SAMPLE>
<IDENTIFIERS>
  <PID >SRR090454</PID>
  <XID ns="Coriell">NA12878 </XID>
</IDENTIFIERS>
```

...

External identifiers must use a namespace (ns) attribute that is registered with the SRA archive.

## Universally unique identifiers

A downstream user of SRA xml data can annotate it with a universally unique identifier. This requires no namespace because it is universally unique (according to the generation method). The INSDC SRAs do not use UUIDs and these are ignored on submission.

```
<RUN alias="68b329da9893e34099c7d8ad5cb9c940" accession="SRR090454" center_name="">
<IDENTIFIERS>
  <PID>SRR090454</PID>
  <UUID> 68b329da9893e34099c7d8ad5cb9c940 </UUID>
</IDENTIFIERS>
```

...